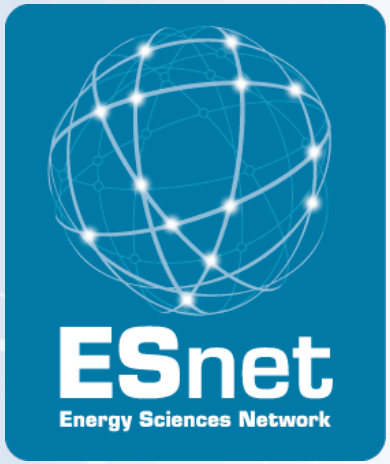


Usage Patterns and Lessons Learned

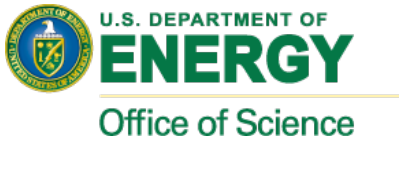
Eli Dart, Network Engineer

ESnet Network Engineering Group

Terabit Networks for Extreme Scale Science Workshop

Washington, DC

February 16, 2011



Overview

Common themes

Discussion of subset of science disciplines that use ESnet

- Examples of trends
- Success stories
- Upcoming needs that are currently unmet

Need for effective tools for extreme scale infrastructure

Common Themes - Science

New science processes such as remote instrument control, experiment health monitoring, etc will place new demands on networks

- Multi-site near-real-time or real-time network interaction
- Need expressed by multiple science communities (light sources, biology, HPC users, etc)
- Many of these communities are not network experts, and will need help from networking organizations in order to progress

Increasing data intensity of science across many disciplines

- Many collaborations that have historically not used the network for data transport must begin soon – 'sneakernet' will no longer be practical
- Many collaborations that have gotten by with using SCP/rsync/etc for WAN transfers will no longer be able to do so – must change to GridFTP or something similar to increase performance
- Collaborations that require >10Gbps connectivity today will need >100Gbps connectivity by 2015 – 10x increase every 4 years

Common Themes - Troubleshooting

Still many performance problems due to packet loss

- Networks aren't clean
- Figure out how to clean them up and keep them clean

Many scientists/collaborations are not network experts, and it is unreasonable to expect them to become experts

- Networks are key to the emerging modes of scientific discovery
  - Widely distributed collaboration
  - Machine-consumable / programmatic interfaces to data
  - Massive data volumes → automated data handling and analysis
- There is a lot of scientific leverage here, but only if the network is an effective scientific tool
- Significant need for middleware and similar tools as networks/systems become larger, more complex, higher performance

High Energy Physics – LHC

Automated data distribution over multiple continents

LHC collaborations are already manipulating large data sets routinely, and the data set size will increase significantly over time

- Nominal data set size today is in the tens of terabytes
- Nominal data set size is expected to be over 100TB by 2014
- 100G trans-Atlantic circuits will be needed by 2014

Service interface to the network

- Automated manipulation of computation and storage today
- Automated scheduling of networks is on the horizon
- HEP science infrastructure operates at very large scale today, enabled by a significant investment in tools and expertise

Nuclear Physics – RHIC at BNL

STAR collaboration

- Widely distributed collaboration
- Significant data transfers to NERSC
- Significant international data transfers
- Over 2PB of raw data to be produced in 2011, 400TB+ of derived data sets to be distributed

PHENIX collaboration

- 2008 data rates from ~650Mbps to 2.4Gbps to Japan (118 TB)
- Near-real-time transfer need (data sets have short lifetime on cache disk, less than 24 hours)
- Data set sizes increasing (projected to transfer 1.2PB to Japan in 2011)

Light and Neutron Sources

ALS at LBL, APS at ANL, LCLS at SLAC, NSLS at BNL, SNS at ORNL, etc.

Large number of beamlines, instruments

- Hundreds to thousands of scientists per facility
- Academia, Government, Industry

Data rates have historically been small

- Hand-carry of data on physical media has been the norm for a very long time: CDs → DVDs → USB drives
- Scientists typically do not use the network for data transfer

Near future: much higher data rates/volumes

- Next round of instrument upgrades will increase data volumes by a factor of 10 to a factor of 100, e.g. from 700GB/day to 70TB/day
- Network-based data transport is going to be necessary for thousands of scientists that will be doing this for the first time in their careers

Light and Neutron Sources

New science architectures coming

- Experiment automation leads to the need for near-real-time health checks
  - Stream sample experiment output to remote location for verification of experiment setup
  - Significant efficiencies of automation are driving this
- Multi-site dependencies (e.g. need for analysis at supercomputer centers)
  - Need a general model for streaming from detectors to supercomputer centers
  - Supercomputer centers often say that allocations change from year to year, therefore significant effort to support one particular scientist may not be wise resource allocation
  - However, many light source users will need to stream data to supercomputer centers – generalized support for this use model will result in significantly increased scientific productivity

Light and Neutron Sources

Some of these data increases have already taken place

Dedicated data transfer hardware and perfSONAR have been used to fix performance problems

- Networks must be loss free
- Networks must be monitored to ensure that they stay clean

These solutions will need to be generalized

- Science DMZs and/or Data Transfer Nodes (DTNs) for light sources
- Assist users with figuring out the “other end” (e.g. suggestions for common architectures such as DTN or Science DMZ)
- Requiring that every collaboration implement their own solution (as many light sources do currently) will result in tens of one-offs over the next few years
  - Difficult to troubleshoot
  - High support load for facility, system and network support staff
  - Therefore, a systematic approach must be developed for large-scale science infrastructure

Climate Science

Supercomputer centers at NERSC and ORNL

Data repositories at LLNL, ORNL, NCAR, NOAA, NCDC, BADC (UK), Germany, Japan, Australia

2PB (~1.6PB and growing) to replicate over multiple continents in 2011

Climate science data sets scale with supercomputer allocations

- Data volume will increase significantly in the coming years
- Data distribution and analysis will result in significantly higher network traffic volumes
- Climate science does not yet have the sophistication of HEP in terms of network-aware tools
  - As their sophistication grows, so will use of the network
  - Automated data distribution is being implemented now

Genomics

JGI does a lot of sequencing and analysis

- Scientists send in samples to be sequenced
- JGI does sequencing/analysis, sends back genome

Significant changes coming

- Price of sequencing equipment dropping dramatically (~$500k to ~ $50k)
- Data rates going up dramatically (1PB/year today, sequencing machine output to go up by as much as 12x over 5 years)

Genomics

Genomics is about to be stood on its head from a data perspective

- Many sites will be able to deploy their own sequencers as costs come down
- Many will not have the local processing/storage infrastructure or systems expertise to do their own analysis
- Instead of sending biological samples to JGI, many sites will send raw sequence data (much larger than the completed genome) to JGI for analysis → significant increase in network traffic

Sites with sequencers can expect larger data flows to/from JGI

This is the beginning – the field of genomics is in its infancy

Fusion Energy – Experiments

Data collection/transfer between EAST in China and DIII-D at GA is a good illustration of the value of dedicated Data Transfer Nodes

- Dedicated servers make data transfer possible
- Two GridFTP boxes built, one shipped to EAST, one deployed at GA
- Data transfers now keep up with experiments

In general, production data transfers are better done on dedicated systems

Fusion Energy – Simulation

Simulations can generate data sets of essentially arbitrary size (e.g. GTC code running at ORNL is expected to generate 500TB/week in a few years)

- Full or reduced data sets must be transferred to other sites (e.g. NERSC, Princeton) for analysis
- Sites that consume fusion data sets from supercomputer centers might want to consider DTNs for their fusion groups if such infrastructure does not currently exist

Fusion Simulation Program (FSP)

- FSP will require integration of multiple sites to run distributed simulation codes

Need for Effective Tools

Many science disciplines are expressing the need for multi-site services

- Unlike HEP and NP, most will not be able to implement these for themselves – BES facilities in particular are going to need a lot of help here
- However, the potential scientific payoff is huge (networks will be able to legitimately claim to have helped enable the next round of breakthroughs in materials, biology, energy efficiency, etc)

Large scale science is inherently multi-site, widely distributed

- As science instruments and supercomputers increase in scale and cost, fewer are built
- The structure of modern science assumes the existence of reliable, high-bandwidth, feature-rich networks to knit the distributed science infrastructure into a coherent whole
- As infrastructure scales up, so does complexity – proper tools are critical

Summary

Many disciplines will need to move large data sets routinely

- Current data transfers use TCP – today, the network must provide loss-free IP service to TCP in the general case (TCP performs poorly with loss, even very small loss rates)
- Many sites and disciplines do not currently have the expertise to manage these data transfers effectively
- New architectures are seeing wider deployment (e.g. dedicated resources for data transfer, both at the network level and the system level)

Multi-site, multi-instrument science will be the norm

- Scientific disciplines are assuming that they can interconnect resources at different sites in an effective and productive way
- Tools to do this are not currently well-deployed at the sites
- Significant need for standardized, widely-deployed tools to enable scientists to integrate multi-site resources

Questions?

Thanks!

Questions?

Thanks!